Harald Lüngen und Marc Kupietz

DAS DEUTSCHE REFERENZKORPUS DEREKO IM JUBILÄUMSJAHR 2014

Die Autoren sind wissenschaftliche Mitarbeiter am Institut für Deutsche Sprache, Mannheim.

Das am IDS seit 1967 aufgebaute Deutsche Referenzkorpus DeReKo (seit 2004 unter diesem Namen) dient der Germanistik als empirische Grundlage für die Erforschung der deutschen Gegenwartssprache. Für Interessierte wurde jüngst in Teubert / Belica (2014) die Geschichte der Korpora und der Korpustechnologie am IDS umfassend nachgezeichnet. Im Jubiläumsjahr 2014 meldet DeReKo nun einen Umfang von über 24 Milliarden Textwörtern, was einem Wachstum von rund 400% gegenüber dem Vorjahr entspricht. Dieser enorme Zuwachs ist u. a. auf zwei große Lizenzabschlüsse im Jahr 2013 zurückzuführen, mit denen digitale Zeitungsdaten und Fachtexte im Umfang von mehr als 16 Milliarden Textwörtern erworben werden konnten, darunter fortlaufende Ausgaben von fast 100 regionalen und landesweit erscheinenden deutschsprachigen Tageszeitungen und Magazinen (z. B. Süddeutsche Zeitung (ab 1993), DIE ZEIT, FOCUS, Neue Zürcher Zeitung, Falter, profil, Weltwoche und Luxemburger Tageblatt (mit Beginn ab 2000 oder später). Damit werden viele Regionen des deutschen Sprachgebiets erstmals oder wesentlich umfangreicher als mit den bisher in DeReKo vorhandenen Zeitungsdaten abgedeckt. Die Karte in Abbildung 1 zeigt die geografische Verteilung und den Umfang von Pressequellen im DeReKo im Jahr 2014; die 2013 schon vorhandenen Quellen sind schwarz, die Neuerwerbungen rot dargestellt. Deutlich zu sehen sind die Zuwächse von Quellen in der Schweiz, in Österreich, in Luxemburg, im Westen und Nordosten Deutschlands, wobei die Abdeckung des Nordens nach wie vor gering ist. Die geografische Distribution war auch ein Kriterium bei der Priorisierung der Quellen für die Integration von DeReKo in das Hauptarchiv von COSMAS II.

IM JUBILÄUMSJAHR 2014 MELDET DEREKO EINEN UMFANG VON ÜBER 24 MILLIARDEN TEXTWÖRTERN.

Anders als die meisten sehr großen (mehr als zehn Milliarden Wörter enthaltenden) Textkorpora verschiedener Sprachen, die derzeit auf internationalen korpuslinguistischen Konferenzen vorgestellt und diskutiert werden (Baroni et al. 2009, Jakubíček et al. 2013, Schäfer / Bildhauer 2012), basiert DeReKo nicht auf einem massenhaften Download und einer massenhaften Aufbereitung von Texten aus dem World Wide Web. Vielmehr legt das IDS Wert darauf, dass die Nutzung von Textdaten in DeReKo rechtlich abgesichert ist. Das beinhaltet in den meisten Fällen, dass mit den Rechteinhabern Vereinbarungen über die sprachwissenschaftliche Nutzung ihrer Daten abgeschlossen werden, was wiederum mit den Urhebern von Webtexten oftmals nicht durchführbar ist. Es gibt aber einen Anteil an Webtexten, der explizit unter einer freien Lizenz veröffentlicht wurde oder für den man von einem Provider umfassende Lizenzen erwerben kann; um solche Texte bemüht sich DeReKo auch. Unter diesen Vorgaben wurden 2011 (in Kooperation mit dem Projekt EuroGr@mm) und 2013 (in Kooperation mit dem Programmbereich Forschungsinfrastrukturen) die Texte der deutschsprachigen Wikipedia-Seiten (Artikel und Diskussionen) als linguistisches Korpus aufbereitet und in DeReKo integriert. Als weitere bemerkenswerte Neuakquisitionen und -aufbereitungen seit 2012 sind das Plenardebattenkorpus PolMine mit Protokollen von Plenardebatten des Bundestags, Bundesrats und aller deutschen Landesparlamente seit dem Jahr 2000 zu nennen, welches im Projekt PolMine von Andreas Blätte (Universität Duisburg-Essen) erstellt wurde, sowie die Akquisition zahlreicher Belletristik-Werke durch neue Lizenzvereinbarungen mit Buchverlagen.

DAS IDS LEGT WERT DARAUF, DASS DIE NUTZUNG VON TEXTDATEN IN DEREKO RECHTLICH ABGESICHERT IST.

Das Projekt Korpusausbau bemüht sich regelmäßig, beispielsweise durch Akquisitionskampagnen und Aktivitäten auf der Frankfurter Buchmesse, um den Erwerb von Rechten für Belletristik-Texte. Es ist aber festzustellen, dass aufgrund der rechtlichen und technischen Situation (Verwertungsrechte der Verlage, Vielfalt der verwendeten digitalen Formate) die Akquisition und anschließende Aufbereitung von Belletristik-Texten pro Wort bis zu

25000 mal aufwendiger (teurer) ist als die von Zeitungstexten. Deswegen ist diese Textsorte nach wie vor unterrepräsentiert in DeReKo. Absolut gesehen sind aber mittlerweile rund 17 Millionen Textwörter an Belletristik enthalten, d. h. durchaus so viel, wie das British National Corpus an ‚imaginative texts' enthält.

Mit dem Umfang von 24 Milliarden Textwörtern ist DeReKo nach wie vor das größte Archiv deutschsprachiger Texte für die sprachwissenschaftliche Nutzung und stellt auch eine Instanz von ‚Big Data' im Sinne einer sehr großen, mit traditionellen Datenbanktechniken nur schwer beherrsch- und analysierbaren Datenmenge dar. Obwohl die heutige schlagwortartige Verwendung des Begriffs ‚Big Data' erst ab 2009 populär wurde (der englischsprachige Wikipedia-Artikel dazu wurde 2010 angelegt), zeigen Kupietz et al. (2014), dass die Textkorpusbestände des IDS, gemessen an den zu früheren Zeiten verfügbaren Speichermedien, sich schon immer, d. h. seit ihren Anfängen unter Paul Grebe und Ulrich Engel im Jahre 1967, in den Dimensionen von ‚Big Data' bewegten. Trotz alledem muss derzeit Speicherplatz nachgerüstet werden, um auch die syntaktischen Annotationen verschiedener Tagger, die normalerweise mit DeReKo ausgeliefert werden, in vollem Umfang bereitstellen zu können.

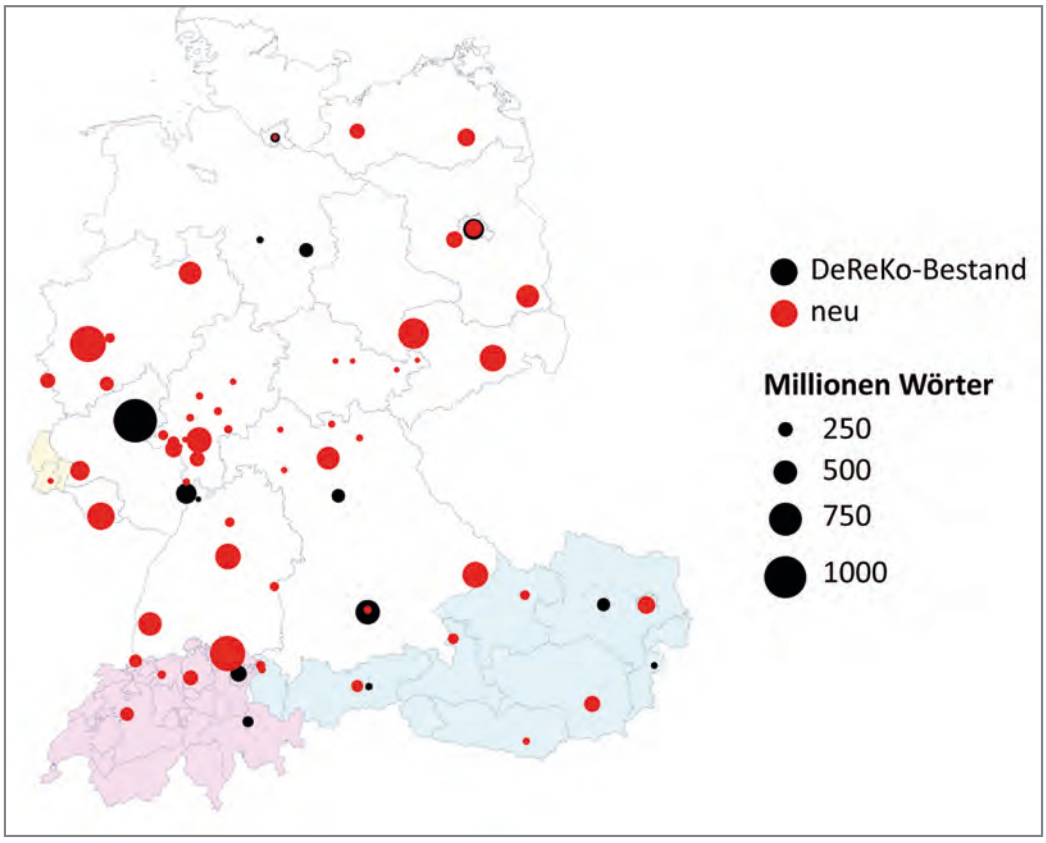

Abb. 1: Karte des dt. Sprachgebietes mit DeReKo-Pressequellen

JE GRÖSSER EIN KORPUS, DESTO ZUVERLÄSSIGER KÖNNEN SCHLUSSFOLGERUNGEN ÜBER SELTENE UND DIVERSIFIZIERTE EREIGNISSE GEZOGEN WERDEN.

Allein durch die gegenwärtig laufenden Nutzungsvereinbarungen beträgt die Wachstumsrate von DeReKo derzeit 1,7 Milliarden Textwörter pro Jahr. Größe und fortlaufende Erweiterung von DeReKo sind aber kein Selbstzweck. Sprache beinhaltet bekanntlich eine große Anzahl seltener Ereignisse – nicht nur lexikalische Ereignisse, sondern auch Ereignisse, die nur durch eine Kombination von Bedingungen beschrieben werden können. Je größer ein Korpus, desto zuverlässiger können Schlussfolgerungen über seltene und diversifizierte Ereignisse gezogen werden. Im Urstichprobendesign von DeReKo (Kupietz et al. 2010) werden zudem virtuelle Korpora anhand von kombinierten Bedingungen über Metadaten definiert (wie „alle Texte des Ressorts Feuilleton in österreichischen Zeitungen der 90er Jahre"), d. h. je größer die Urstichprobe, desto größer und damit repräsentativer für eine Forschungsfrage können auch die virtuellen Subkorpora ausfallen. Daher gilt auch für DeReKo weiterhin das Motto, dass mehr Daten bessere Daten sind (vgl. Church / Mercer 1993). DeReKo dankt allen seinen mittlerweile fast 200 Textgebern, ohne die das Deutsche Referenzkorpus nicht möglich wäre.

Abb. 2: DeReKo Textgebergrafik, Stand 2012

Literatur

Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros (2009): The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In: Language Resources and Evaluation 43(3), S. 209-226.

Church, Kenneth / Mercer, Robert (1993): Introduction to the special issue on computational linguistics using large corpora. In: Computational Linguistics 19 (1), S. 1-24.

Jakubíček, Miloš / Kilgarriff, Adam / Kovář, Vojtěch / Rychlý, Pavel / Suchomel, Vít (2013): The TenTen corpus family. In: Abstract Book of the 7th International Corpus Linguistics Conference CL2013. Lancaster University, S. 125-127.

Kupietz, Marc (2014): Der Programmbereich Korpuslinguistik am IDS: Gegenwart und Zukunft. In: Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim: Institut für Deutsche Sprache, S. 320-328.

Kupietz, Marc / Belica, Cyril/Keibel, Holger / Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC'10). Istanbul: ELRA, S. 1848-1854.

Kupietz, Marc / Lüngen, Harald / Bański, Piotr / Belica, Cyril (2014): Maximizing the Potential of Very Large Corpora. In: Kupietz et al. (Hg.): Proceedings of the LREC-2014-workshop Challenges in the Management of Large Corpora (CMLC2). Reykjavik: ELRA, S. 1-6.

Kupietz, Marc / Lüngen, Harald (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik: ELRA, S. 2378-2385.

Schäfer, Roland / Bildhauer, Felix (2012): Building large corpora from the web using a new efficient tool chain. In: Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12). Istanbul: ELRA, S. 486-493.

Teubert, Wolfgang / Belica, Cyril (2014): Von der linguistischen Datenverarbeitung am IDS zur „Mannheimer Schule der Korpuslinguistik". In: Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim: Institut für Deutsche Sprache, S. 298-319.